7. SPRHW -- Science Processing HWCI

The Science Processing HWCI (SPRHW) is the primary HWCI in the Data Processing Subsystem and contains processing resources (processors, memory, disk storage, and input/out subsystems) necessary to perform first time processing, reprocessing, and Algorithm Integration & Test (AI&T). SPRHW also provides the hardware resources necessary to support management of the science processing, in the form of a Queuing Server and Production Planner Stations. However, the queue management function in science processing is closely coupled to the ECS planning function, and the Queuing Server and Production Planner Stations are therefore discussed in the Planning Subsystem Design Specification, 305-CD-026. This chapter describes only the hardware resources required to execute the science software.

The design specification for SPRHW is derived by first reviewing the requirements for SPRHW in Section 7.1. The requirements are traced from the first principles established in the system requirements, through the detailed processing requirements levied by the instrument teams, and through the models built to simulate system behavior under the load defined by the instrument teams. The technology assessments performed to support SPRHW specification are then reviewed in Section 7.2. The conclusions of the technology assessments are then applied to the derived, detailed system requirements to produce the SPRHW specification, provided in Section 7.3. Details of the specification are then discussed in Section 7.4 to show how they meet the system requirements.

This subsystem specification provides a general discussion that is applicable to each DAAC. The requirements for SPRHW, and hence the detailed SPRHW specifications, are very DAAC specific; for these details, the reader is referred to the Design Specifications for the DAACs:

A. GSFC: DID-CD-305-030-002;

B. LaRC: DID-CD-305-031-002;

C. EDC: DID-CD-305-033-002;

D. NSIDC: DID-CD-305-035-002; and

E. JPL: DID-CD-305-036-002.

Note that ASF and ORNL do not perform science processing within ECS; therefore, these DAACs do not have SPRHW components.

7.1 SPRHW Requirements Analysis

The specification of SPRHW must satisfy sizing, scalability, reliability, maintainability, availability, compatibility, and interoperability requirements. These are discussed in the sections below.

7.1.1 SPRHW Sizing and Expandability

The SPRHW sizing and expandability requirements are derived from several sets of fundamental system requirements established in the ECS Functional and Performance Requirements Specification (F&PRS). Specific numerical detail is provided for these requirements by the processing plans provided by the instrument team inputs to the Ad Hoc Working Group on Production (AHWGP). Because the processing plans are quite complex, with many inter-relationships between products, significant analysis is required to understand the processing plans in order to reach a processing hardware design. This analysis has been performed using static and dynamic modeling. The modeling and design efforts have identified some design parameters that are not addressed by the AHWGP inputs; to address these design parameters, additional information concerning the processing plans has been solicited from the instrument teams in the form of a survey. The information derived at each of these steps has been used to determine the design parameters (requirements) for the science processing hardware. The following sections describe that process.

7.1.1.1 Functional and Performance Requirements

The F&PRS provides several requirements specifying the throughput and timeliness of ECS science processing.

Requirements EOSD1050, EOSD1060, EOSD1070, and EOSD1080 state the timeliness requirements for producing products for first time processing of data. Level 1, Level 2, and Level 3 products must be produced with in 24 hours of the time that ECS receives all of the data required to produce the product. Level 4 products must be produced with seven days of the time that ECS receives all of the data required to produce the product.

Requirement EOSD1040 states that ECS shall have the capacity to perform reprocessing at twice the rate of first time processing. Requirement PGS-1300 states that ECS shall have the capacity to perform AI&T, production of prototype products, ad hoc processing for "dynamic browse" or new search and access techniques, and additional loads due to spacecraft overlap at one times the rate of first time processing.

Requirement PGS-1301 states the requirement that vendor peak processing estimates for processors be derated by a factor of four for design purposes.

Requirement PGS-1270 states the requirement that the processing design and implementation have expandability by a factor of up to three without design change, and by a factor of up to ten without major design change.

7.1.1.2 The Technical Baseline and the AHWGP Inputs

Appendix C of the F&PRS is included by reference in the F&PRS requirements. This appendix contains estimates of the volume of processing to be performed by ECS. Because these estimates have changed over time and become significantly more detailed as instrument team plans have matured, the material in Appendix C of the F&PRS has been replaced by a group of separately maintained documents referred to as the ECS Technical Baseline.

The Technical Baseline contains material that defines the level of processing required by each instrument for first time processing. Much of this data has been provided by the instrument teams in the form of inputs to the AHWGP. These inputs are updated approximately every six months; the most recent updates to the AHWGP inputs and the ECS Technical Baseline were baselined in February, 1996. The February 1996 Technical Baseline is therefore the basis for the CDR design for SPRHW.

AHWGP processing requirements are expressed in terms of Product Generation Executives (PGEs), which are the smallest scheduled increment of science processing. For each PGE, the AHWGP input identifies the DAAC at which the PGE will be executed, and provides the execution frequency, the list of input data requirements, the list of output data products, and an estimate of the number of millions of floating point instructions (MFPI) performed with each execution. For each input file, an estimate is given of the fraction of the file that is read by the PGE. The baseline also identifies the calendar quarters during which the PGE will be executed; this allows the instrument teams to identify how they will phase in their processing after the instruments are put on orbit. As a result, the processing demand for each instrument may vary by calendar quarter.

The Technical Baseline documents program plans and directives that affect the SPRHW specification. The hours of operation for each DAAC are provided. For DAACs having less than 24 hour per day, seven day per week operations, the assumption is made that all processing must be accomplished during DAAC operating hours.

The Technical Baseline also documents the program directive to phase the implementation of processing capacity for each instrument relative to the date of the instrument's launch. The purpose of the phasing factors is to provide sufficient early processing resources to support AI&T, without purchasing the required full-up capacity before it is needed. With **X** defined as the processing resources required to do first time processing for an instrument and **L** defined as the launch date of the instrument, the phasing factors are defined below:

- **0.3X for L-2 < t < L-1.** Pre-launch AI&T requires 0.3X during the period from one to two years before launch.
- **1.2X for L-1 < t < L+1.** Pre-launch AI&T and system I&T requires 1.2X during the year before launch. Standard instrument processing requirements (X) begin from the launch date and last for the remainder of the life of the instrument.
- **2.2X for L+1 < t < L+2.** Post-launch AI&T, standard processing, and reprocessing of data require 2.2X starting at launch plus one year.
- **4.2X for t > L+2.** Post-launch AI&T, standard processing, and reprocessing of data require 4.2X starting at launch plus two years.

The launch dates for instruments supported by ECS are documented in the Technical Baseline. Because the PGE processing plans are expressed on a quarterly basis, for the purpose of applying the phasing factors the launch dates are brought forward to the first day of the appropriate calendar quarter. The large majority of processing requirements supported by Release B are for instruments on two platforms, TRMM and AM-1. TRMM is scheduled for launch during the third quarter of (calendar year) 1997 (3Q97), and AM-1 is scheduled for launch during the third quarter of (calendar year) 1998 (3Q98). Therefore the phasing factors for SPRHW are assessed at 3Q96, 3Q97, 3Q98, 3Q99, and 3Q00. However, Release A of ECS has already established resources to cover the 3Q96 requirements; hence the focus of Release B design is for the latter four dates.

Specifications for SPRHW have been designed for each of the four Release B dates identified above. However, for the purpose of discussing the design, the 3Q99 period (also referred to within ECS as Epoch K) has been selected as a common frame of reference across the entire set of Release B Design Specifications. At Epoch K, TRMM instruments are supported at the 4.2X level of processing, and AM-1 instruments are supported at the 2.2X level of processing. For the sake of simplicity, and because their processing requirements are relatively small, the COLOR, SWS, DFA, MR, and SAGE III instruments are phased as if they were launched simultaneously with AM-1. The phasing requirements for DAO are not related to instrument launches, and are documented separately in the Technical Baseline.

7.1.1.3 Static Modeling

The AHWGP inputs define over two hundred PGEs, executed at five DAACS for eleven instrument teams. The PGEs create over three hundred products, and require over one hundred ancillary and Level 0 input data sets. The interdependencies of the products and PGEs creates a network so complex that it cannot be understood by hand. The inputs are therefore analyzed using models, to reduce the volume of data into aggregates for each instrument and DAAC, and to understand the dynamic behavior of the interdependencies. The first step in this analysis is to use a static model (spreadsheet) to analyze the PGEs.

The static model represents each PGE within ECS on a single line of a spreadsheet. The columns of the spreadsheet contain the attributes for the PGE: its process identifier, instrument, processing site, frequency of initiation, number of executions per initiation, total input file size, total output file size, number of input files, number of output files, volume of files staged per execution, volume of files destaged per execution, and number of MFPI per execution.

Most PGEs are executed once per set of inputs, and those inputs are produced (by the instrument and/or by other PGEs) periodically. However, some PGEs may be executed several times on a given set of inputs that are periodically produced. An example is a MODIS PGE that produces products defined in terms of geographic tiles; for each set of input data covering the whole earth, the PGE may be executed 355 times, once for each MODIS tile. Hence it is necessary to track both the frequency of initiation of the PGE (the periodicity of the inputs) and the number of times the PGE is executed for a given number of inputs.

The numbers and sizes of input and output files are evaluated by summing the inputs provided for the PGE by the instrument team. The staging and destaging inputs to the static model are derived from the AHWGP inputs. Staging refers to the process of moving data to the science processing resources, and destaging refers to moving data away from the science processing resources. SPRHW does not archive data; it retains over extended periods (more than 24 hours) only a very limited amount of ancillary data identified by the instrument teams as permanent. All other data required by or produced by processing must be staged or destaged to the ECS archive. The staging and destaging entries in the static model spreadsheet are calculated by subtracting the permanent ancillary file volumes from the total file volumes.

These inputs are then used to calculate average resource usage levels for processing, network input/output (I/O), archive I/O, and disk I/O. The processing usage is calculated as the number of millions of floating point operations per second (MF) required to meet the PGE's requirements based on the PGE's frequency of initiation, number of executions per initiation, and MFPIs per execution. Similarly, the data flow to the archive (due to destaging), from the archive (due to staging), over the network (due to staging and destaging), and over the disk I/O channels (due to processing, staging, and destaging) are calculated. These values are then aggregated (summed) for each instrument and DAAC.

One method of static model analysis is to calculate the average level of resource usage required to support each PGE, instrument, and DAAC, assuming that the processing is spread smoothly over time. This analysis ignores the timeliness requirements for products, but it produces a baseline estimate of the minimum resources required to keep up with the processing requirements. It is this analysis that defines the value of X used for phasing requirements.

Another approach for analysis is to manipulate the formulas used to calculate the derived values (the usage rates) to take into account the timeliness requirements; further, it is assumed that in the worst case each periodic product is executed at least once on the worst case day. This produces a (generally) worst case estimate of the resources required to meet the timeliness requirements for first time processing. (Because the static model cannot take into account the dynamic interactions between PGEs, it is theoretically possible that the worst case for the system is actually exceeded by this analysis, but the dynamic model has not shown this to be the case.)

Static modeling results for each DAAC are presented in the DAAC Design Specifications.

7.1.1.4 Dynamic Modeling

The static model has significant limitations. It cannot take into account the dynamic interdependencies of the PGEs or the system's computing resources. As a result, it cannot accurately predict the end to end clock time required to produce a stream of products once Level 0 data has been received. It also provides extremely inflated estimates of the system staging requirements, because it must assume that every (non-permanent) file required for each execution of each PGE is staged for that execution; it cannot take into account that the file may already have been staged to SPRHW in support of some other PGE. On the other hand, it assumes uniform utilization of the SPRHW resources, not accounting for the fact that one type of resource (such as a processor) may be idle while the system uses another type of resource (a network).

To overcome these limitations, a dynamic model of a subset of ECS has been implemented. The dynamic model is implemented using the Block Oriented Network Simulation (BONeS) tool. BONes is a discrete-event simulation tool for analysis and design of communication networks and distributed processing systems. Components of a distributed processing system are represented by nodes, which have resources associated with them that get allocated as events request them. First time production of products by DPS is simulated by the Processing module, in conjunction with the Data Handler module, the Event Driven Scheduler, and the Ingest module. The Data Handler simulates the behavior of the ECS Data Server. The Data Handler stores and retrieves data from the permanent archive, routes data to the requesting subsystems, and manages tiered storage and staging resources. The Event Driven Scheduler monitors the availability of data, requests data to be staged from the Data Handler to Processing, routes newly created data to the appropriate Data Handler or Processing resource, and initiates execution of a process when all required inputs are present. It should be noted that the Event Driven Scheduler is not intended to serve as a simulator of the Planning Subsystem; rather, it simulates the effects of data arrival at ECS and the actions of the Planning Subsystem. The Ingest module simulates the Ingest subsystem, which accepts data from external systems and users and contains rolling storage of Level 0 instrument data.

The dynamic model is run by creating an input event stream representing the arrival times of Level 0 and ancillary data files. The PGEs are defined to the model in terms of their resource requirements and file dependencies. The system design is also represented as an input to the model, in terms of the number and capacity of system resources -- for example, the number of SPRHW processors and their throughput in MF. The Event Driven Scheduler then executes the event stream, which is driven by the external events (receipt of Level 0 and ancillary files) and by derived events (the availability of the data and computing resources required to execute a PGE). The simulation is run over an extended simulation period -- typically 21 days -- to examine the effects of products produced periodically. The simulation is started with an empty ECS system; there is therefore a starting transient in the simulation's output results. Because of this, the results of the first five days of the simulation are generally discarded.

The results of the dynamic model are analyzed to identify peak and average resource requirements, and average and maximum system response characteristics. Graphs showing resource utilization over time during the period of the simulation can also be generated.

The model explicitly represents each processor within SPRHW, and constrains the simultaneous execution of PGEs to the number of processors in the design, operating at the vendor's peak MF rating derated by a factor of four per the requirements. Average and peak processor utilization and average and peak processor queue lengths are reported, as well as processor utilization curves over the period of the simulation. These results identify whether processing backlogs are occurring, and whether there is excess processing capacity in the design.

For disk storage at the science processors, the model does not constrain the available space; instead, it implements an algorithm to depict the purging of unneeded data from the SPRHW storage array, and monitors the demand for storage. By plotting the storage demand over the period of the simulation, it can be determined when the peak occurs, and for what duration. This information is used to size the SPRHW storage arrays.

Network and disk I/O are averaged over time to provide average utilization rates. The model represents these as constrained resources; hence, the model results provide assurance that a given level of network and disk I/O throughput will satisfy the system's requirements, and provide an approximation of the fraction of these resources that is used on average.

Dynamic modeling results for each DAAC are presented in the DAAC Design Specifications.

7.1.1.5 Memory Usage and I/O Survey

The inputs to the AHWGP do not provide any information as to how the PGEs will use processing memory, or about the detailed characteristics of their disk I/O.

Although all of the potential platforms considered for SPRHW make use of virtual memory, the ratio of physical memory (RAM) to virtual memory required during the operation of a system is important for performance reasons. The rate at which memory is accessed by a PGE and the pattern of these accesses can also have important performance impacts. For the SPRHW specification, these requirements affect the amount of RAM, number of memory channels, amount of disk (for virtual memory), and number of disk channels (for virtual memory access). Because the SPRHW processing requirements are very demanding and generally require sophisticated and expensive processors, it is desirable to get as much throughput as possible from the processors by keeping as much of a PGE in RAM as practical.

Similarly, the size and pattern of disk I/O requests strongly affects the performance that can be expected from a disk I/O subsystem. Because of various overheads associated with each disk access, throughput from a disk subsystem increases strongly with the average size of the disk request, up to maximum limits imposed by the disk, interface, and controller technologies. While vendors usually quote I/O subsystem performance for large request sizes (hence peak performance values), it is important to know whether these request sizes are typical for ECS applications; and if not, to derate the vendor performance specifications accordingly.

The type of memory usage and file I/O characteristics that are needed to determine the system requirements for these resources can generally be determined either by code inspection or by measurement during execution. ECS solicited this information from the Release B instrument teams in January of 1996. Unfortunately, it is still too early for most teams to provide accurate assessments of their requirements, and little data has been received to date in response to the survey. The integration and test of beta science software for Ir-1 during the first half of 1996 may provide requirements in time to support the procurement of the first increment of Release B hardware.

For purposes of the generation of the Release B CDR design, assumptions have been made in the absence of requirements data in these areas. It is assumed that each PGE will require approximately 128 megabytes (MB) of RAM. This figure is viewed as a reasonable upper limit for most algorithms, especially if the algorithms are designed with consideration of their memory requirements. It is recognized that some algorithms may require large data structures to iteratively refine their data using models, and will have correspondingly larger memory requirements. These uneven needs between algorithms are mitigated in part by the Symmetric Multiprocessor (SMP) architecture that has been selected for SPRHW; this is discussed in Section 7.4.

It is also assumed that ECS SPRHW file I/O will be buffered by the platform operating system, with the average request size characterized by the operating system I/O buffer size. This assumption is made because it is believed that this correctly characterizes the use of HDF, which is expected to be the dominant form of disk access on the science processors. The buffer size varies by operating system, but is typically small (64 kilobytes or less) relative to request sizes used for direct file I/O (256 KB to 4 MB). This means that the vendor disk I/O performance ratings will be derated to reflect small request sizes.

Where RAM information estimates are available from the instrument teams, the DAAC Design Specifications describe the information provided and its impact on the design for that DAAC.

7.1.2 SPRHW Reliability, Maintainability, and Availability Requirements

The product generation function of DPS has an availability requirement of 0.96 (96%) and a mean down time requirement of not to exceed four hours. Implicit in these requirements is also the requirement that SPRHW be able to achieve it timeliness and throughput requirements while allowing for 4% non-availability and component outages of up to four hours.

7.1.3 SPRHW Compatibility and Interoperability Requirements

The Release B ECS software must satisfy ECS requirements for the use of open systems technology, and must be compatible with other Release B subsystems and with hardware provided to the DAACs for Release A. Specifically, the SPRHW hardware suite must be supported by

OODCE. SPRHW must also support the network interfaces to be implemented by ECS. For the DAACs with smaller ECS requirements, these interfaces are FDDI; for the DAACs with larger ECS requirements, these interfaces must support data exchange between DPS and DSS at aggregate rates (in Epoch O) of on the order of 150 MB per second.

7.2 Technology Assessments

The requirements identified in Section 7.1 can be compared to the capabilities of current and future technologies to identify candidate technical solutions. This type of evaluation can be done on paper, in the form of trade studies or technology evaluations, or it can be done in the laboratory, through prototyping and benchmarking. Sections 7.2.1 and 7.2.2 below discuss the technology assessment activities that have been performed to support the Release B design process, with emphasis on the activities performed since IDR-B.

7.2.1 Technology Evaluations

A number of technology evaluations have been performed under ECS to identify the best technical approaches and products for the SPRHW suite. Three of these evaluations are discussed below.

7.2.1.1 Production Platform Families

The requirements for first time processing for instruments at the DAACs range from a few MF for the less compute intensive instruments, up to 4.5 GF for MODIS at GSFC. Computing requirements in the GF range have historically been the turf for vector supercomputers, but advances in massively parallel processors, symmetric multiprocessors, and workstation farms have pushed these technologies into this range of performance. An evaluation was performed for Release A and Release B of the following candidate platform family architectures for SPRHW:

A. Single processor workstations;

B. Farms of workstations connected via local area networks;

C. Symmetric Multiprocessors (SMPs);

D. Massively Parallel Processors (MPPS); and

E. Vector Supercomputers.

The key criteria used to evaluate the platform families were lifecycle cost per unit of processing power, scalability, ease of science software development, and flexibility. The evaluation found that the requirements for the less compute-intensive instruments could be met by single processor workstations, which are the least expensive and most flexible resources. For the compute-intensive instruments, however, only the SMPs, MPPs, and vector supercomputers offered the scalability required. Among these three, the vector supercomputers were found to have significant cost disadvantages. The SMPs were found to have significant advantages over the MPPs in the ease with which applications can be developed, and the evaluation concluded that the SMP platform class was best suited for ECS's high-end SPRHW requirements.

This evaluation is documented in *Platform Families for the ECS Project*, 440-TP-007-01.

7.2.1.2 Distributed and Parallel Processing

Parallel processing in both shared memory and distributed memory architectures is viewed as an emerging technique for solving large processing problems. An analysis has been performed to examine the benefits of distributed and parallel computing for ECS. The analysis studies various processing alternatives for ECS science algorithms and provides information on processing technologies. The analysis examines the applicability of using the OSF/Distributed Computing Environment (DCE), symmetric multiprocessing with shared memory, distributed multiprocessing (including workstation clusters), and massively parallel processing for ECS science software.

The SMP architecture is found to provide significant flexibility: it fully accomodates non-parallelized code, supports easy migration of non-parallel code to shared memory parallel processing, and can also support distributed memory parallel processing. This flexibility is viewed as giving SMP systems a significant advantage for ECS processing. Although a need to parallelize particular PGEs has not been identified to date, the ability in the future to parallelize ECS science software is an important consideration, especially with the large processing requirements of Release B and beyond.

This evaluation is documented in *Distributed and Parallel Processing For ECS Science Algorithms: A Trade Analysis*, 440-TP-008-01. ECS Science and Technology Lab prototyping efforts that have demonstrated parallel program development, based primarily on parallelization tools, are documented in 194-00569TPW and 194-430-TPW-001.

7.2.1.3 Production Topologies

A technical analysis has been performed examining the advantages and disadvantages of distributing processing tasks from one or more instruments across one or more processing clusters. The purpose of the study was to examine how the interactions between the processing requirements of different instruments might increase or decrease the total hardware resources required to perform the aggregate load. The results of this analysis are reflected in the SPRHW design. By DAAC, the conclusions of the analysis are as follows:

A. LaRC. The compute-intensive instruments at this DAAC (CERES TRMM, CERES AM-1, and MISR) were found to be large enough to require whole platforms to meet their requirements. It will be beneficial to cross over the execution of certain CERES TRMM monthly processes to the CERES AM-1 processing suite, because this allows the two sets of monthly processes to share some data sets, and reduces the load on the LaRC archive. No similar synergy between CERES and MISR was found.

B. EDC. Some potential exists for ASTER to benefit from unused processing cycles on the platform specified for MODIS at EDC, because the MODIS processing is strongly cyclic, and the average utilization on this platform on some days of the MODIS cycle is low. However, beginning at Epoch K (3Q99), MODIS reprocessing will absorb almost all of this idle capacity.

C. GSFC. The requirements for LIS and COLOR are small and of comparable size, and are easily fit to a common host. These instruments are not rolled onto the large MODIS platforms at GSFC because any future variations in MODIS requirements might be larger than the base requirements for LIS and COLOR, making the planning for these instruments problematic.

D. JPL. The requirements for SWS and DFA/MR are small and of comparable size, and are easily fit to a common host.

E. NSIDC. NSIDC supports processing for only one instrument.

This analysis is documented in *Production Topologies: A Trade-Off Study Analysis for the ECS Project*, 440-TP-006-001.

7.2.2 Prototypes and Benchmarks

Prototypes and benchmarks are used to assess the applicability of a technology or product to a set of requirements. Two sets of recently acquired benchmarks have been used in refining the ECS SPRHW design for Release B. The first set of benchmarks, measuring the performance of IP over HiPPI, was performed in the ECS development facility. The second set of benchmarks, measuring I/O subsystem performance on the SGI platforms, were performed at the University of Minnesota and by SGI.

7.2.2.1 Network Benchmarking

Benchmarking tests of the performance of IP over HiPPi have been performed in the ECS development facility. The purpose of these tests was to measure the throughput that can be achieved using IP over HiPPI, to determine whether IP introduces a significant overhead in HiPPI communications.

The tests used two SGI Challenge XL systems, equipped SGI HIO HiPPI adapters. These systems were running version 5.3 of the Irix operating system. Two different benchmarking tools were used to send TCP streams from the memory of one system to the memory of the second system. Throughput rates were measured as the configuration parameters for TCP/IP were varied.

The maximum transfer rates observed in the test were approximately 55 MB per second, just over half of the 100 MB per second theoretically possible with HiPPI. SGI has stated that using version 6.2 of Irix, rates of up to 90 MB per second have been observed in their laboratories. This indicates that IP does not burden HiPPI with a substantial overhead, and that IP-based protocols can be used over HiPPI in ECS. This eliminates the need to develop custom protocols for ECS using raw HiPPI interfaces.

7.2.2.2 I/O Subsystem Benchmarking

The Laboratory for Computer Science and Engineering (LCSE) at the University of Minnesota, in affiliation with the Army High Performance Computing Research Center (AHPCRC), has performed a number of benchmarking tests on I/O subsystem performance on SGI platforms. These tests have included performing high speed data transfers over HiPPI using a lightweight TCP/IP protocol, and performing high data rate disk transfers. These tests have demonstrated that the SGI architecture can sustain I/O rates in excess of 150 MB per second between machines, and over 500 MB per second to locally attached file systems.

The HiPPI experiments performed at LCSE demonstrated the ability of the SGI platforms to support data visualization. In these experiments, a single Challenge server, configured with four HiPPI connections, was used to transfer data from its file system to the graphics processors of two SGI Onyx systems. A lightweight TCP/IP protocol referred to as "NFS-Bypass" was used to perform the transfer. This protocol, which provides an NFS-like interface, transfers data between systems using a socket to socket connection, bypassing most of the overhead normally imposed by NFS. Transfer rates for a single HiPPI connection, from disk to graphics processor, of 65 MB per second have been observed. Using multiple HiPPI connections transfer rates from a single machine of over 150 MB per second have been observed.

Research at LCSE has also been performed to evaluate the maximum transfer rates available from various storage configurations on SGI platforms. These experiments have focused on the use of striped, SCSI-2 based RAID file systems using SGI's XFS file system and direct I/O. The LCSE researchers have built file systems capable of sustaining over 500 MB per second transfer rates on a Challenge system. A key aspect of their findings in the building of fast file systems is the need to use large block request sizes. They found that block request sizes associated with buffered I/O -- I/O buffered by the operating system's I/O cache -- are 64 KB or less. Disk subsystem performance at these request sizes is substantially below peak levels; at this request size, throughput of less than 10 MB per second is observed on SCSI-2 channels having a theoretical bandwidth of about 20 MB per second. Using direct I/O, which is not buffered by the operating system's cache, and with request sizes of one MB or more, throughput rates rapidly approach the per channel limitation imposed by SCSI-2.

SGI has further tuned the NFS Bypass software to make efficient use of the Challenge memory subsystem, and is productizing the software. The product, Bulk Data Service (BDS), will be offered as an extension to Irix 6.2 in the second quarter of 1996. Significantly, since BDS is built on top of TCP/IP, it is not specific to HiPPI; BDS may be used to implement highspeed file transfers over any TCP/IP connection.

SGI has also recently performed benchmarking tests on their newest generation of RAID controllers. These tests indicated sustained transfer rates of eight MB per second per controller are achieved when small request sizes (64 KB) are used. This benchmark is used as a basis for sizing the disk I/O subsystems for SPRHW in Release B.

7.3 SPRHW Specification

SPRHW has three top-level component types: Science Processors, Queuing Servers, and Production Planner Stations. As discussed earlier, the Queuing Server and Production Planner Station components are specified in the Planning Subsystem Design Specification.

The Science Processors are characterized by their vendor, enclosure, processors, memory, I/O subsystems, internal disk drives, external disk systems, backup and update devices, network interfaces, and monitors. This section provides a description of the selected hardware and is as specific as possible. It does not enumerate numbers of components, however, as this level of detail is specific to the DAACs. For this level of detail, please refer to the DAAC Design Specifications.Each of these is discussed below.

7.3.1 SPRHW Vendor

The Release B SPRHW components will be provided by Silicon Graphics (SGI). The architecture trade-offs show that SMP systems provide clear advantages for ECS science processing, and SGI is a leading vendor of SMP systems. When considerations are taken into account for cost, availability of required ECS software capabilities (such as OODCE), availability of highspeed network components (particularly HiPPI), and ease of re-use of Release A hardware, the logical choice for Release B hardware is SGI.

There are two sets of components for which final selections have not been made: the SPRHW suite to support DAO, and the external disk systems for SPRHW at all DAACs. DAO SPRHW requirements are different from other ECS SPRHW requirements because DAO is actively pursuing distributed memory processing techniques in order to scale to grand challenge processing requirements. DAO is prototyping distributed memory science algorithms and evaluating vendor hardware and software, and will make its hardware selections based upon these tests. The procurement and implementation of DAO resources is following a different schedule than that of the rest of ECS, which is driven by the launch of the TRMM and AM-1 platforms.

As for the external disk systems required by SPRHW, a baseline design is presented here which features RAID-5 devices using SCSI-2 interfaces. This technology is mature and its performance is fairly well understood. However, new technologies, particularly fibre channel based RAID arrays, are making their way into the market. If these products are found to be stable and cost effective before the first Release B procurement, they may be selected as the Release B baseline.

7.3.2 Enclosures

The actual enclosures for the science processors and their disk arrays are generally of little interest, except that they impose constraints on expansion capabilities, and they occupy floor space at the DAACs. There are four types of science processor enclosures included in the SPRHW specification: desktop workstations, tower workstations, Challenge and Power Challenge DM and L enclosures, and Power Challenge XL enclosures.

The desktop workstations and tower workstations are used in small DAACs, where processing requirements are low. They generally have very little internal expansion capacity in terms number of processors or internal disk drives. They can be rack mounted, although the desktop workstations are generally configured with high-power graphics capabilities and are assigned a dual role for Algorithm Quality Assurance; these system are expected to be placed in the ECS DAAC operations area.

The Challenge and Power Challenge DM and L configurations use the same enclosure, a half height cabinet. ECS Operations and Maintenance is investigating the extent to which these units can be rack mounted vertically, to save floor space at the DAACs. In terms of expansion capabilities, the most important limitation is that these systems offer only five backplane slots; Section 7.4.1 discusses expansion options for these units.

The Power Challenge XL uses a full height cabinet. This enclosure offers fifteen backplane slots, and can be expanded to offer as many 25 backplane slots. The expansion options for this unit are discussed in Section 7.4.1.

The enclosures for the external disk storage units are discussed in Section 7.3.1.N.

7.3.3 Processors

For new equipment purchased for Release B, SPRHW will use the MIPS R10000 processor. This chip has a superscalar 64 bit architecture, capable of performing two floating point operations per clock cycle. SGI has announced that the chip will be offered in a 275 MHz implementation in the second half of 1996. Using the derating factor defined in the F&PRS, this processor is estimated to provide 137.5 MF; the dynamic modeling runs made to support Release B CDR have been based on this processor.

The R10000 will be offered across the SGI product line, from workstations through compute servers. In the DM enclosure, configurations of one, two, or four processors are possible. In the L and XL configurations, the processor is offered on boards having either two or four processors; hence, the L and XL systems can be configured with any even number of processors, up to their backplane limits.

7.3.4 Memory

In Section 7.1 above it was noted that data for memory requirements are not available yet for many instruments, and thus it would be assumed that each processor required approximately 128 MB of RAM.

For single processor workstations, a configuration with 128 MB of RAM has been specified.

For SMP configurations, the SGI memory architecture features leaf controllers, which process requests to a subset of the system memory. Each memory board can have one or two leaf controllers. A system can have up to eight memory boards (subject to backplane limitations), but only up to eight memory controllers. The I/O benchmark results by the group at the University of Minnesota have shown that memory interleaving is important for I/O performance. It is also expected to be important for science algorithm performance on systems with many processors. Therefore, machines with more than one GB of RAM and machines performing high rates of I/O will be configured with eight memory controllers; lesser machines will generally be configured with four memory controllers.

SGI offers RAM at two chip densities. Combined with options for up to eight memory boards, there are a large number of memory configurations offered. Unfortunately, however, there is a gap in the offerings with eight way interleaving between two GB and four GB. As a result, systems with between 14 and 24 processors are configured with two GB of RAM, and systems with more than 24 processors are configured with four GB of RAM.

7.3.5 I/O Subsystems

The workstation science processors do not offer configurable I/O subsystems; I/O interface cards (device control cards and network interface cards) plug directly into the backplane. The number of backplane slots in these systems is sufficient to support the device and network cards required for the ECS configurations, and the I/O rates for these systems are low.

For the SMP systems, the I/O rates range up to 100 MB per second, and the number of devices required to achieve these rates is large. The SGI architecture provides configurable I/O subsystems which attach to the backplane; each subsystem occupies one backplane slot and provides up to 320 MB per second of bandwidth. The subsystem card is referred to as a PowerChannel 2 or IO4 card. A system may have up to six IO4 cards, subject to backplane limitations.

The first IO4 card in an SMP system provides console and ethernet connections. Each IO4 provides serial and parallel connections, two fast wide differential (FWD) SCSI-2 channels, a space for a VME controller, and space for two HIO controller cards. HIO controller card offerings include a HiPPI card, a FDDI card, and a card supporting three SCSI-2 channels.

The number of IO4 cards specified for each SMP is determined by allocating HIO slots to the FDDI and HiPPI interfaces (if required), and counting the number of SCSI-2 interfaces required. The number of SCSI-2 interfaces required is determined by the number of internal and external SCSI-2 devices supported by the system. In general, it is assumed that the internal slow SCSI-2 devices (CD-ROMs, floppy disk drives, and tape drives) will be aggregated on the first SCSI-2 channel. Internal disk drives will be allocated to the second SCSI-2 channel. External disk arrays will be allocated to subsequent SCSI-2 channels; the number of channels required is based on the desired throughput of the file system, and is discussed below in Section 7.3.1.N.

7.3.6 Internal Disk Drives

The bulk of storage for the science processors will be provided by external storage arrays. Internal disk drives will only be used to provide swap space for the operating system, and to provide file system space for the operating system and applications. Applications in this context is not intended to include the PGE executables or any temporary file space required by the PGEs; rather it refers to file system space requirements for the Autosys client software and ECS custom code. The file system requirement for the operating system and applications is not expected to exceed two GB.

The allocation for swap space is estimated at four times the size of the physical memory (RAM). This exceeds what is generally configured for servers, and probably represents an upper bound; if it is found that systems require virtual memory larger than four times their physical memory, it is likely that physical memory will also have to be upgraded to reduce paging.

SGI currently offers 2 GB and 4.3 GB internal disk drives and will soon offer 9 GB internal disk drives. A combination of these drives have been selected for each system to satisfy the estimated space requirement. The number of internal disk drives in these combinations ranges from one to four. These drives will be allocated to a single SCSI-2 channel.

7.3.7 External Disk Arrays

The disk size and throughput requirements for SPRHW are determined on a DAAC by DAAC and host by host basis, using the dynamic modeling results. These requirements are translated into numbers of drives and controllers based upon the configurations available from the vendor. The current product of choice for external disk arrays is the RAID-5 product from SGI.

The SGI RAID-5 arrays use one redundancy disk for each four data disks. Arrays are built in groups of five disks, with up to four groups (20 disks) in an enclosure. Up to four enclosures can be put in a rack.

An enclosure will support one or two controllers. Each controller can access one or more groups of disks in the enclosure. A group of disks within an enclosure can be accessed by both controllers; however, only one controller may access the group at a time. This form of dual attachment is useful only for implementing failover of the controllers (or hosts) without moving cables. Each controller within an enclosure can also be tied to more than one host; however, only one host may mount the associated file system at a time, and therefore this form of dual attachment is also only useful to implement failover. The controllers in an enclosure can be connected to different hosts, and in fact this configuration is specified for some DAACs.

The throughput to the SGI arrays is limited by the either the interface mechanism (at large request sizes) or by overheads in the host and the controller (at small request sizes). The SCSI-2 interface used in the SGI arrays allows a maximum bandwidth of 20 MB per second to a channel. At small request sizes, however, the overheads are expected to limit performance to not more than eight MB per second per controller. In order to avoid contention on the SCSI-2 channels that would significantly reduce throughput for large block requests, and would also modestly impact throughput for small block requests, each array controller is configured on its own SCSI-2 channel.

The number of SCSI-2 channels and array controllers is determined by dividing the throughput rate required for a system (determined from the static and dynamic modeling results) by the eight MB per second rate per channel assumed for small request sizes. This number is then rounded upward. (Although the disk I/O performed by a system will actual be a mixture of large and small block requests, no effort is made to determine an average throughput rate; rather, the slower rate associated with small block requests is used to represent the entire load. This increases the design margin for these components.) The number of enclosures and racks is then matched to this number of enclosures.

SGI offers 2 GB and 4.3 GB disks in its arrays and soon will also offer a 9 GB disk. The 2 GB disks are not specified for use in SPRHW because the SPRHW disk requirements are too large to make use of these drives practical.

Although the number of groups and size of disks within groups can be varied from one controller to another, the number of such variations assigned to a single host has been deliberately limited. This is becasue it is expected that file systems will be defined so that they are striped across multiple controllers. Such striped file systems are necessary to support very high speed devices such as the HiPPI network. Striped file systems can only be built using partitions of uniform size, which are most easily built when each controller within the stripe is controlling the same configuration of disks.

In some cases, the need for controllers is greater than the need for disk space, and enclosures and racks are specified that are only partially full. This is unavoidable because the number of controllers per enclosure can be increased beyond two.

The disk arrays will be configured with three power supplies (one redundant supply) and a minimum amount of write cache. (Vendor benchmarks suggest that write cache will not significantly improve performance for the request mix that SPRHW is expected to produce.)

7.3.8 Backup and Update Devices

Each science processor will be equipped with an internal CD-ROM, as this is the delivery mechanism employed by SGI for updates to operating system software. Workstations will also include a floppy drive, part of their default configuration.

Each SPRHW suite will include at least one machine equipped with a tape library. These tape libraries will provide backup capability for SPRHW and for other subsystems, using the FDDI backbone network. For Ir-1 and Release A, Exabyte EXB-210 and EXB-218 tape libraries were purchased and put in place at LaRC, GSFC, and EDC. The tape library requirements for JPL and NSIDC are to be determined; our working assumption is that they can be satisfied by a single 8 mm DAT tape library. The EXB-210 is an 8 mm Digital Audio Tape (DAT) library with eleven cartridge slots and two drives. The EXB-218 is a 4 mm DAT with 19 cartridge slots, two operational drives, and one hot spare drive.

The tape drives will generally be configured to machines designated for AI&T use.

7.3.9 Monitors

In general, the science processors are to function as compute servers, and are not configured with sophisticated graphics processing capabilities. It is expected that the machines will be housed in raised floor environments where space is at a premium. The science processing hosts require a console in order to monitor the system at startup, and to perform some system administration tasks; however, this console need not be a bulky graphics monitor. ECS Operations and Maintenance are investigating ways to rack the system consoles, and to allow systems to share the same console.

Where desktop workstations are specified for use at a DAAC, the purpose is to share a processing resource between SPRHW and the Algorithm Quality Assurance HWCI (AQAHW). For these machines, a sophisticated graphics capability is provided, and a full size graphics monitor is required. It is expected that these machines will be housed within the DAAC operations area, to support their AQA operations.

7.3.10 Network Interfaces

A FDDI subnetwork will be implemented at each DAAC to support the Planning and Data Processing Subsystems, with the exception of ASF and ORNL, which do not implement the ECS Planning and Data Processing Subsystems. Each processing component of SPRHW (including the Queuing Server and the Production Planner Stations) will be interfaced to the PDPS FDDI subnetwork; the design of this subnetwork is specified in the Design Specification Overview, CD-305-020, and in the DAAC Design Specifications. For the Challenge and Power Challenge SGI science processors, this interface is implemented with an HIO mezzanine card, occupying one slot on a PowerChannel 2 (IO4) board. SGI Challenges and Power Challenges networked to FDDI prior to Release B were interfaced using a VME board connected to the IO4 board; these interfaces will be left in place in Release B. For SGI workstations included in SPRHW, the FDDI interface is implemented with a GIO card, which occupies a slot on the system motherboard.

At LaRC, EDC, and GSFC, the data transfer requirements between DPS and DSS necessitate the implementation of a switched HiPPI network. The HiPPI network will be implemented via a central HiPPI switch with switched 800 Mbps interface ports connected directly to the SPRHW and DSS hosts; this design is also documented in the Design Specification Overview and the DAAC Design Specifications. For the Challenge and Power Challenge SGI science processors, this interface is implemented with an HIO mezzanine card, occupying one slot on an IO4 board.

7.4 SPRHW Design Discussion

Section 7.4 provides a discussion of how the specification provided in Section 7.3 meets the requirements identified in Section 7.1. Most of this discussion is specific to particular DAAC configurations, and is provided in detail in the DAAC Design Specifications. The sections below provide a general discussion of how the specifications meet the sizing and expandability requirements, and what the general SPRHW failure recovery strategy will be.

7.4.1 Sizing and Expandability

It is important to note that although the science processors are divided into clusters of resources configured to deal with specific processing requirements, and are expected to be tuned and allocated for these purposes, this does not imply that the use of a processing cluster, or a processor within that cluster, is limited to one operating mode or instrument. The DAAC operations staff, using the planning tools provided by ECS, will determine how the DAAC work load is distributed across the pool of SPRHW resources at the DAAC.

The DAAC Design Specifications contain tables showing that the systems specified for the DAAC meet the requirements for the DAAC, as determined from the modeling results. Inspection of these tables shows that at each epoch, sufficient processing resources are specified to perform the level of processing anticipated for that period.

The ability to expand the SPRHW configurations merits a separate discussion. A DAAC's science processing capabilities can be expanded in two ways: by adding resources to existing hosts and disk arrays, and by adding new hosts and disk arrays. The former approach is referred to as upgrading in place.

The ability to upgrade in place is limited by the number of additional resources that can be added to a host. For the workstation science processors, this is pretty much limited to memory upgrades and possible future processor upgrades. Memory configurations of up to 640 MB are available on these systems.

For the SMP hosts, the ability to upgrade in place is limited by the number of backplane slots that are unused in the base configurations, and by artificial limits that are built into the SGI product line.

The Challenge DM is limited to one processor board (up to four processors) and a combination of four memory or IO4 boards. Since at least one IO4 board is required, the maximum memory configuration would feature three boards, or six GB of memory with six-way interleaving. If only one memory board were used, three IO4 boards could be used. This would support six HIO cards, plus six SCSI-2 channels. This combination could be used to support a FDDI interface, a HiPPI interface, and enough SCSI-2 interfaces (a total of 18) to support a 100 MB per second file system. The Challenge DM can also be upgraded in place to a Power Challenge L or Challenge L.

The Power Challenge L and the Challenge L have five backplane slots, and their use is not artificially constrained, although every system must have at least one processor board, one memory board, and one IO4 board. A system may therrefore have up to twelve processors, or the memory and I/O configurations described above for the Challenge DM.

Most of the machines called out in the DAAC Design Specifications are Power Challenge XLs. The Power Challenge XL has 15 backplane slots, although this limit can be expanded to 25 by using a backplane extender. The maximum number of processors supported by Irix 6.2 on the Power Challenge is 36; this would require nine processor boards. Up to eight memory boards can be used, although only eight leaf controllers can be configured on those boards. Up to six IO4 boards can be put in the system. Thus a maximally configured Power Challenge XL would require 23 of the 25 available slots. Such a system would provide 4.95 derated GF, with 16 GB of RAM. The I/O subsystem would support multiple HiPPI connections and SCSI-2 controllers. (In benchmarking experiments at the University of Minnesota, Challenge systems have been configured with as many as four HiPPI interfaces and as many as 40 SCSI-2 controllers -- but not simultaneously.)

The largest (fullest) systems called out in the Release B SPRHW specification are designed to meet the compute-intensive requirements of MODIS at GSFC. These machines are configured with 24 processors, two GB of RAM, and three IO4 boards supporting 14 SCSI-2 channels, a FDDI interface, and a HIPPI interface. By upgrading in place, the processor capability of this system could grow by 50 percent (from 24 to 36), the memory could grow by 700 percent (from two GB to 16 GB), and the number of disk channels and network interfaces could be doubled.

Upgrading in place will therefore satisfy potential future requirements to grow processor throughput by 50 percent or more, memory requirements by 700 percent or more, and disk and network I/O by as much as 100 percent.

If growth beyond these scales is required, it will be necessary to add more hosts to the SPRHW configuration. The HiPPI switch supports 16 ports, but if necessary the switch can be cascaded; adding a second switch would thus provide 30 ports for DPS and DSS hosts. If the ratio of DPS hosts to DSS hosts is assumed to be two to one, this would allow connections for 20 DPS hosts, or a maximum SPRHW compute power of 99 GF (derated). This is roughly 400% of the 25.25 GF specified for LaRC at Epoch O, which has the highest aggregate compute requirements.

If processing power greater than 99 GF (derated) is required, it would be necessary to modify the design slightly by adding a second HiPPI network, and partitioning the HiPPI network traffic. (Cascading the HiPPI network top three switches is not recommended by the vendor, although it may be feasible.)

It should be noted that adding SPRHW hosts potentially has the effect of increasing network traffic, because the SPRHW storage subsystems are locally (host) attached, and can not be shared between hosts. As more hosts are added to process the same load, the potential that data will have to be replicated to multiple host file systems increases, and the aggregate file system size, disk I/O, and network I/O also increase.

7.4.2 Failure Recovery Strategy

The general strategy in the event of the failure of an SPRHW component is to re-distribute the first time processing load to SPRHW components having sufficient resources to handle the load. Reprocessing and AI&T, although important to ECS over an extended period of time, can be delayed or reduced until the failed components are replaced. Because SPRHW has redundant capacity to support AI&T and reprocessing, and because scheduling of SPRHW resources is performed dynamically by the queuing and planning resources, most failures will have little impact on the timeliness of production processing.

7.4.2.1 SPRHW Failure Recovery

The DPS reliability, maintainability, and availability (RMA) requirements of 96% availability and mean down time of less than four hours are met by the SPRHW design. An analysis demonstrating this is provided in *Availability Model/Predictions for the ECS Project*, 515-CD-001-003.

In the event of a failure in the SPRHW suite, the DAAC operators will isolate the fault and determine its severity. Vendor maintenance will be called and repair and/or replacement of the affected part will be initiated.

Depending on the severity and location of the failure, other management actions and decisions may be required. If the failed component only affects resources that were allocated to AI&T, there is no impact to DAAC production operations, and no further action to recover production activity is required.

If the failed component affects resources that were allocated to production, an assessment will be made whether any PGEs need to be re-started or re-run because of possible corruption. If this is the case, ECS production planning capabilities will be used to reschedule the PGEs. If the failed component has affected a limited subset of the production resources, the MSS subsystem will notify the Queuing Server that these resources are no longer available, and the Queuing server will appropriately stop scheduling these resources. If the remaining resources available for production processing are sufficient to keep up with the first time processing plan, an immediate re-plan may not be necessary; if the first time processing PGEs are prioritized above reprocessing PGEs (generally the case), the first time processing will remain on schedule. If it is necessary to shut down SPRHW resources in order to facilitate a repair, the a re-plan can be generated, allocating time in the schedule for the repair.

If a failure reduces the working SPRHW resources allocated to production below the level required to support first time processing, the DAAC operators may choose to re-allocate AI&T resources to production. This may be done in one of two ways: by using the production and queuing software to logically re-allocate and re-plan the use of the system's resources, or by using the physical hardware in the AI&T hosts to replace failed hardware in the production hosts. The first approach, using the ECS system software to schedule production work on the AI&T platforms, is generally preferred, because it is not intrusive to the hardware. However, there may be circumstances when the second approach is simpler and faster. If the AI&T software environment is significantly different than the production environment (suppose AI&T was testing a new version of an operating system), and if the hardware failure in production is simple (a failed power supply), then changing hardware may be easier than changing software.

Data backup and recovery are a comparatively minor issue for SPRHW, because SPRHW does not provide long term, secure storage for data. Although SPRHW has large storage arrays, these arrays are used to hold ancillary data files and data granules for short periods of time. If a file system failure occurs within SPRHW, the algorithms, ancillary data files, and data granules can be recovered from the Data Server. Therefore the only data that would be backed up from SPRHW, and restored in the event of a failure, would be operating system and configuration files (i.e., the system disk). In Release B, MSS will provide hardware to perform backup and recovery over the FDDI backbone. In the event that recovery is needed, incremental and full backup tapes would be used to re-build the required file systems.

7.4.2.2 Network Failure Recovery

The FDDI subnetwork for the Planning and Data Processing Subsystems provides a significant degree of fault tolerance in the physical communications system. Most media failures within the FDDI fabric will not result in any loss of service and no reconfiguration would be necessary in these cases (due to the basic nature of FDDI). Given the inherent fault tolerance of FDDI, it is not required to have multiple physical communications paths to each host. Hosts within SPRHW will use single-attached station cards.

Failure recovery for the HiPPI switch will be supported by stocking spares for the Line Replacable Units of the switch (power supplies, interface cards, fan). Individual interface cards can be changed without disrupting other hosts. If the control module fails, it will be swapped out and the switch will be reconfigured. In the very rare event of an entire switch chassis failure, the switch would either be replaced or repaired. All these failure recoveries involve activities on the switch; modifications to the attached hosts are generally not required.